EXHIBIT E

Hidden complexity of free energy surfaces for peptide (protein) folding

Sergei V. Krivov* and Martin Karplus*†‡

*Laboratoire de Chimie Biophysique, Institut de Science et d'Ingénierie Supramoléculaires, Université Louis Pasteur, 67000 Strasbourg, France; and †Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138

Contributed by Martin Karplus, August 24, 2004

An understanding of the thermodynamics and kinetics of protein folding requires a knowledge of the free energy surface governing the motion of the polypeptide chain. Because of the many degrees of freedom involved, surfaces projected on only one or two progress variables are generally used in descriptions of the folding reaction. Such projections result in relatively smooth surfaces, but they could mask the complexity of the unprojected surface. Here we introduce an approach to determine the actual (unprojected) free energy surface and apply it to the second β-hairpin of protein G, which has been used as a model system for protein folding. The surface is represented by a disconnectivity graph calculated from a long equilibrium folding-unfolding trajectory. The denatured state is found to have multiple low free energy basins. Nevertheless, the peptide shows exponential kinetics in folding to the native basin. Projected surfaces obtained from the present analysis have a simple form in agreement with other studies of the β-hairpin. The hidden complexity found for the β-hairpin surface suggests that the standard funnel picture of protein folding should be revisited.

For relatively rigid systems (e.g., many organic molecules), and for flexible systems with a small number of significant degrees of freedom (e.g., short peptides), it is now possible to determine the potential energy surface and free energy surface (FES) by sampling the minima and saddles and constructing disconnectivity graphs to describe them (1–5). Direct extension of these methods to proteins (polypeptide chains with many degrees of freedom, a well defined native state, and a denatured state with a large number of conformations) has not been possible. Thus, most approaches used for such systems have resorted to essential simplifications in their description of the potential energy surface and FES. Progress variables, usually involving one or two degrees of freedom [e.g., number of native contacts, number of H bonds, number of native dihedral angles, rms deviation (rmsd), radius of gyration], have been selected and the FES was determined as a function of these variables. The projected FESs have generally been found to be relatively simple with a single or several low free energy barriers (a few kT or less). These results support the concept that a smooth FES provides the bias necessary for folding. Although the popular "funnel picture" is phrased in terms of the energy (6, 7), it is the FES that determines the folding behavior. In many proteins (8) the loss of entropy on folding nearly counterbalances the effective energy, and calculations for some model systems suggest that the entropy loss introduces an activation barrier, that leads to two-state folding (9). This result makes it all of the more important to determine the complexity of the FES of peptides and proteins. The well characterized β-hairpin of protein G (10) is studied here by a recently developed approach (4) based on disconnectivity graphs (1).

An unprojected representation of the FES of a protein can be obtained and presented in a meaningful way by using an equilibrium trajectory to construct a transition disconnectivity graph (TRDG) (4). The essential idea (see *Methods*) is to group the states into free energy minima, not according to their geometrical characteristics (such as the number of native contacts) but rather according to the equilibrium dynamics; i.e., from an equilibrium trajectory we determine the populations of the states, which provide the relative free energies, and the rates of the transition between the states, which provide the free energy barriers. A schematic example is shown in Fig. 1*a*; the corresponding TRDG is shown in Fig. 1*b*. The FES consists of three basins, within which there are many transitions and they are connected by transition state ensembles (TSEs), through which pass relatively few transitions. The number of times a system is found in each basin gives an estimate of the partition function of the basin, and the number of transitions between two basins gives an estimate of the partition function for the TSE between the basins. When there are multiple paths between basins, the TSE can be obtained with the Ford–Fulkerson theorem for determining the maximum flow (reaction rate) between the nodes (FE minima) of a network (4). For the case of the folding-unfolding transition of a protein this method needs to be extended to avoid trivial (uninteresting) partitions of the denatured basin. We briefly describe an approach, called the balanced minimum cut method, to solve this problem for the general folding-unfolding case.

The analysis tools we developed are used here to study the β-hairpin of protein G, which NMR has shown to fold in the absence of the rest of the protein to a rather well defined state (10). This system is of interest because it is small enough to be studied by the present approach and, despite its small size, has been proposed as a model system for protein folding based on experimental studies (11, 12). In particular, the measurements have been interpreted in terms of a two-state folding transition, like that observed in many small proteins.

Methods

System. The β-hairpin used for this study has the sequence Gly-Glu-Trp-Thr-Tyr-Asp-Asp-Ala-Thr-Lys-Thr-Phe-Thr-Val-Thr-Glu with no blocking groups for the terminal residues; the model is the same as that used in ref. 13 for a Monte Carlo (MC) study. The peptide was modeled by the CHARMM program (14) with the polar hydrogen potential function (15) and the EEF1 implicit solvation model (16). The results of explicit solvent simulations (17) are in general agreement with those obtained with EEF1, but use of the latter makes possible a much more complete sampling of the configuration space. The β-hairpin was simulated at 360 K for 4 μs to obtain a sufficient number of folding/unfolding events for analysis. Structures were recorded every 20 ps. This process provides the data necessary to obtain the coarse-grained properties of the important portion of the FES. The statistics concerned with the largest basins and the transitions between them, which are of primary interest and likely to be robust, are accumulated much more rapidly than the fine-grained properties, i.e., the number of times the system passes through the TSE between two multiconfigurational basins

Abbreviations: FES, free energy surface; rmsd, rms deviation; TRDG, transition disconnectivity graph; TSE, transition state ensemble; MD, molecular dynamics; MC, Monte Carlo.

‡To whom correspondence should be addressed. E-mail: marci@tammy.harvard.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0406234101

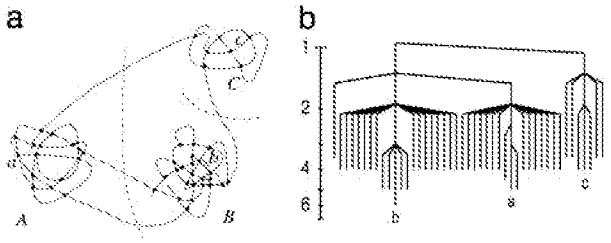

Fig. 1. Schematic model system showing a trajectory and the corresponding TRDG. (*a*) Three FE basins (A–C) are shown with the transitions between them. Dots show clusters that were visited more the once and the connecting line corresponds to a single transition found in the trajectory. Dashed lines show the minimum cuts that separates A from B and C and B from C. (*b*) The TRDG was constructed as described and shows that the FES consists of three basins; the lowest (representative) nodes in each basin are labeled as in *a*.

is accumulated much faster than that for a particular configuration of the TSE.

Clustering. The 2×10^5 structures obtained from the trajectory were divided into 35,377 clusters and 83,331 transitions (i.e., different entries in the transition matrix). The clusters were defined by using an all-atom rmsd cutoff of 2.0 Å (corresponding to a backbone rmsd of ≈1 Å). Each subsequent structure was compared with the set of clusters found so far; if the rmsd of the structure from the first configuration of all of the known clusters exceeded a given threshold, it was taken as a new cluster. The element n_{ij} of transition matrix $\{T\}$ equals to the number of transitions from cluster j to cluster i found in the simulation. The matrix $\{T\}$ is such that MC kinetics with transition probabilities $p_{ji} = n_{ji}/\Sigma_j n_{ji}$ reproduces the actual kinetics obtained from the trajectory. When clusters of configurations are used as states, as is necessary for large systems like proteins and even for the β-hairpin, this equivalence does not follow automatically and may depend on the clustering method; e.g., use of too large an rmsd cutoff or use of secondary structure clustering significantly lowers the barrier between the denatured and native basins of the β-hairpin. The clustering procedure and rmsd used here were verified by showing that they yield the same estimates for the partition functions of the native and denatured states, and for the folding time, as the original molecular dynamics (MD) trajectory (see below). The rmsd clustering with the subset of coordinates used by Dinner *et al.* (13) gives very similar results, indicating that the analysis is robust.

Partition Functions. The partition function of state i is taken to be $Z_i = \Sigma_j n_{ij}$, the number of times the system visited the state. The undirected graph with edge capacities $c_{ij} = (n_{ij} + n_{ji})/2$, is used to represent the kinetics at equilibrium (4). The partition function of the FE barrier separating states i and j, Z_{ij}, is equal to the value of minimum cut between the states in the graph, which can be calculated by the Ford–Fulkerson algorithm (18). Z_{ij} is close to (it is an upper bound, since there may be recrossings) the number of transitions (direct and indirect) between the states i and j. After calculating the minimum cuts (FE barriers) between every pair of nodes (clusters), which can be done with only n-1 total minimum cuts for n nodes by use of the Gomory–Hu algorithm (19), the TRDG (4) is constructed to obtain a detailed representation of the FES. Following Becker and Karplus (1) and taking into account that $F_{ij} \sim -kT \ln(Z_{ij})$, one starts with the largest Z_{ij} (smallest F_{ij}) and successively connects states in order of decreasing Z_{ij} (increasing F_{ij}).

Minimum Cut and Balanced Minimum Cut Methods. To separate states (clusters) of two basins, say A and B in Fig. 1*a*, we have to find a node in basin A such that a minimum cut between that node and a corresponding node in B separates A from B. Nodes a in A and b in B are such nodes, since the minimum cut (shown as a dashed line in Fig. 1*a*) that separates a from b also separates A and B; we call a and b representative nodes (usually these are the most visited nodes in a basin). However, a basin may not have such a node; e.g., an example is the entropic basin found here. In such cases the balanced minimum cut procedure introduces an "extra" node that is connected to all nodes in the graph with the same small capacity c and is used as a representative node for the denatured basin. Illustrative examples of the minimum cut and balanced minimum cut procedure are given in *Supporting Text* and Figs. 7–9, which are published as supporting information on the PNAS web site.

Folding Kinetics. To find the distribution of first passage times to the native structure, we performed an MC simulation with transition probabilities p_{ij} found from transition matrix $\{T\}$ (see above). A total of 10,000 trajectories were started from the first cluster (corresponding to the completely extended structure). These results are compared with 35 folding trajectories calculated by MD starting with the completely extended structure.

Results

The β-hairpin from protein G (10) (see *Methods*) was selected for study because many simulations (13, 20–22) and some experimental folding data (11, 12) are available. Moreover, activated dynamics (21) and transition path sampling (22), as well as projected FESs (13, 20), have suggested that the folding of this peptide may have some inherent complexity. Interestingly, a recent paper (23) on the β-hairpin with a specialized potential function has obtained a reptation-like folding path not found in earlier studies.

The β-hairpin was modeled with the CHARMM program (14) and the EEF1 force field (16). The TRDG graph obtained from the 4-μs trajectory at 360 K is shown in Fig. 2. One can clearly distinguish four major FE basins plus many minor basins; i.e., the FES does not have a single funnel structure. Although the native basin, by itself, is funnel-like, the denatured basin is not. The lowest free energy basin (the native state at T = 360 K) corresponds to a β-hairpin, but it is different from that found at 300 K (13); basin 1 (Fig. 2) is the second lowest basin at 300 K (13). This result is not surprising since the free energy of folding calculated at 300 K (with the same potential function) is only −0.4 kcal/mol (13) and structure 1 is somewhat less tightly packed, so that its basin has a larger entropy than the lowest state at 300 K. This result emphasizes the importance of the entropy in determining the "native" state in this marginally stable model for the β-hairpin peptide at different temperatures.

The denatured state can be seen from Fig. 2 to consist of basins 2–4 and a very large number of clusters that make up an "entropic" basin. The entropic basin could not be separated into subbasins; i.e., all partitions were closely connected with small equilibration times (1 ns or less). Fig. 3 shows some representative low free energy structures in the denatured basin.

The TRDG in Fig. 2 shows explicitly that the FES, including the denatured state, does not have a simple (funnel) structure. However, we note that, although the free energy determines the folding behavior, the usual funnel diagram plots an effective energy [i.e., the potential energy plus the solvation free energy (6, 7)]. Correspondingly we examine the effective energy, $< U >$, of the five major basins; the thermodynamic properties are given in Table 1. Basins 2–4 are lower in effective energy than basin 1, so that their higher free energy is caused by their narrowness (lower entropy), relative to the basin 1. From the form of the TRDG and the $< U >$ values, the potential energy surface, like the FES, has a complexity that is in disagreement with the simple funnel picture (6, 11, 24).

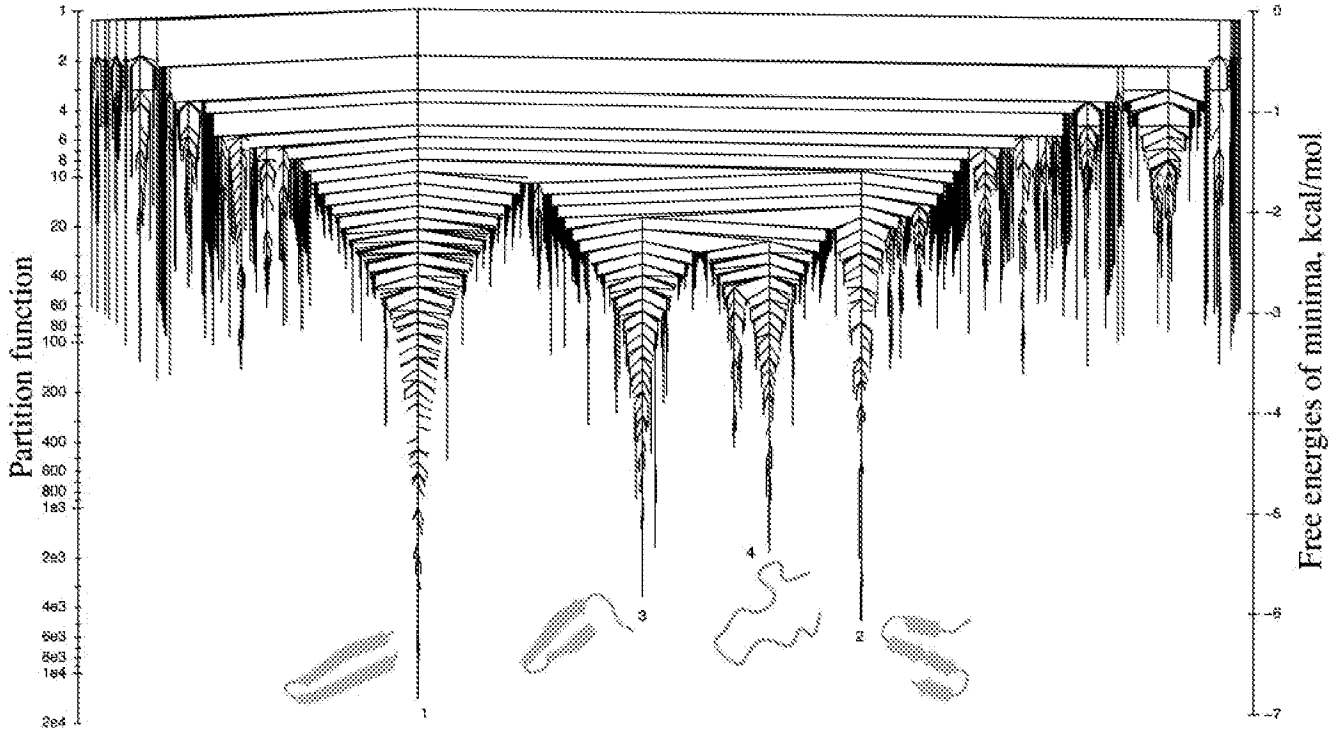

Fig. 2. TRDG of β-hairpin calculated with EEF1 solvation model (16) at 360 K. Representative structures for the deepest FE minima are shown. The left vertical axis shows the Z_i for the minima and the Z_{ij} for the barriers. The right vertical axis shows $F_i = -kT\ln(Z_i)$ and $-kT\ln(Z_{ij})$ in units of kcal/mol for the minima and barriers, respectively. The free energy barrier is $F_{ij} = -kT\ln(Z_{ij}) - kT\ln(Z_{ij} \times h/kT \times 1/t_q) = -kT\ln(Z_{ij}) + 3.61$ (4, 33); h is Plank constant and $t_q = 20$ ps is the sampling interval. For example, the height of the free energy barrier between the native basin and the denatured basin is about $-kT\ln(8 \times 6.29.10^{-3}) + kT\ln(69{,}065) =$ 10 kcal/mol, and the mean unfolding time is 69,065/8 $\times$ 20 ps $\approx$ 173 ns, where 69,065 is the partition function of the native basin (Fig. 5).

To determine whether the complex TRDG found here is consistent with published studies, which show relatively smooth free energy landscape for the β-hairpin (13, 20–22), we used the data from the folding/unfolding trajectory to obtain projected results. Projection of the FES shown in Fig. 2 onto the rmsd from the NMR structure and radius of gyration as progress variables yields Fig. 4*a*, which is similar in form to those found in refs. 13 and 17 and for some proteins (25). As already noted, the lowest free energy state at 360 K (labeled 1 in Fig. 2) is not the native state at 360 K, so its rmsd from the NMR structure is rather large (4 Å); the secondary minimum (rmsd $\approx$ 8 Å) seen in the projected surface corresponds to basin 2 of Fig. 2. The contrast between Figs. 2 and 4 is striking and provides a cautionary insight concerning the interpretation of such surfaces. Projection on two principal components (3) also reduces the complex TRDG to a simple funnel-like surface (see Fig. 4*b*); this result is similar to that observed for Ala_{12} in ref 26.

Because of the complex form of the FES found for the β-hairpin, it is helpful to describe the folding behavior by constructing a small network (Fig. 5), which is based on the TRDG and preserves the minimum cuts (i.e., the free energy barriers) between the basins. During folding, the denatured system spends considerable time (77% of the time spent in the entire denatured basin) in basins 2–4, before entering native basin 1. Although basins 2–4 are "intermediate" states, they are "off-pathway," since the probability to go from any of them directly to basin 1 is essentially zero and the system has to pass through the entropic basin, which serves as the dynamic "hub" for the folding reaction. Correspondingly, most transitions between the important basins of the denatured state go through the entropic basin. Interestingly, a recent study of the fast folding villin headpiece subdomain (27) suggests that the denatured state has basins with well defined structures.

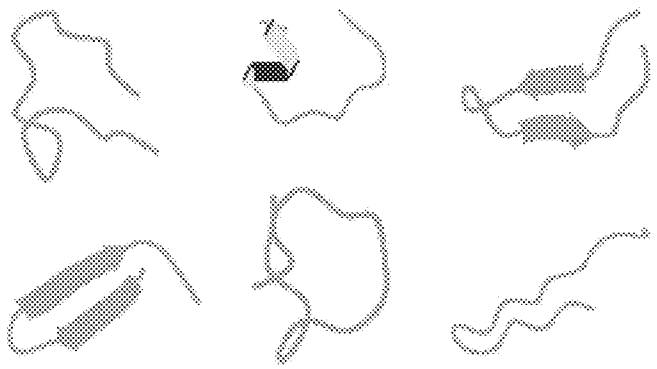

Fig. 3. Structures of the most-visited clusters in the entropic basin of the denatured state.

Table 1. Properties of the native basin, the three well defined denatured basins, and the broad entropic basin; see Figs. 2 and 4

No.	F	$\langle U \rangle$	TS
1	0	−348.91	0
2	0.72	−351.79	−3.6
3	0.51	−350.28	−1.9
4	0.36	−349.59	−1
Entropic	0.59	−345.02	3.30

$F = -kT\ln(Z)$ is free energy, $\langle U \rangle$ is mean potential energy calculated over all clusters in the basin (determined by the minimum cut and balanced minimum cut procedures), and $TS = F - \langle U \rangle$. F and TS are given with respect to the lowest free energy basin. Units are kcal/mol.

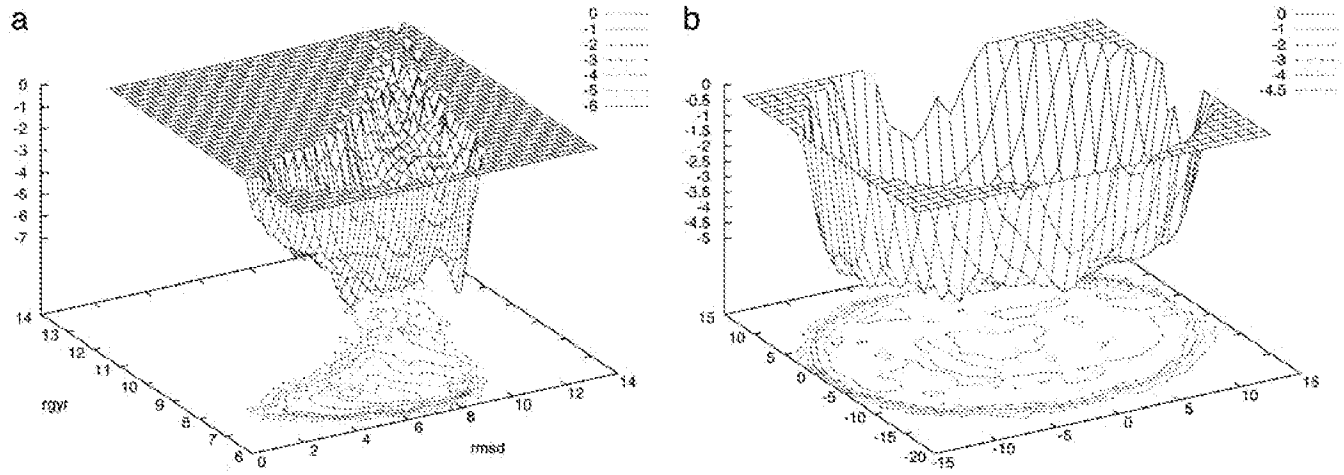

Fig. 4. FESs (kcal/mol) projected on two dimensions. (*a*) The rmsd from NMR native structure and radius of gyration in Å. (*b*) The two most important principal components in Å.

To explore the folding kinetics, a number of trajectories starting with the completely extended structure state were run and the first passage time was determined (Fig. 6). Good agreement is obtained between the MD folding simulation and MC folding times (see *Methods*). It demonstrates that the latter preserves the kinetics, validating the clustering procedure and indicating that the 4-μs trajectory is reasonably converged. The folding kinetics shows single exponential behavior within the accuracy of calculations. This result, which is in agreement with experimental analyses (11), is of considerable interest because of the multiminimum nature of the FES of the denatured state. As indicated by Fig. 5, the equilibrium within the multibasin denatured state is sufficiently fast that its complexity does not significantly alter the single exponential character of the folding time distribution.

The calculated folding time of the β-hairpin is relatively fast; i.e., it is $\approx$350 ns from the MD trajectory and the time estimated from the transition matrix, with the clusters used in constructing the TRDG, is similar (422 ns). The calculated folding times are faster than the experimental estimate of 6 μs at room temperature (11), presumably because of the neglect of solvent friction and the higher temperature used for the trajectories.

Concluding Discussion

The approach presented here has made possible the determination of the unbiased (unprojected) FES for the folding of a model for the β-hairpin of protein G. We recognize that the potential function, with a continuum representation of the solvent, is approximate and that the quantitative results we report may not correspond exactly to those for the physical system studied experimentally. Nevertheless, the qualitative features, particularly the hidden complexity of the FES, are expected to be meaningful.

In a recent paper Evans and Wales (28) have also studied the FES of the β-hairpin with the same potential and used selected folding paths [the discrete path sampling method (29)] to draw a free energy disconnectivity graph. They obtained a simple funnel surface, essentially corresponding to that of our native state basin. The denatured surface is not sampled because the discrete path sampling method is designed to find the fastest path between two given structures, rather than to provide a proper

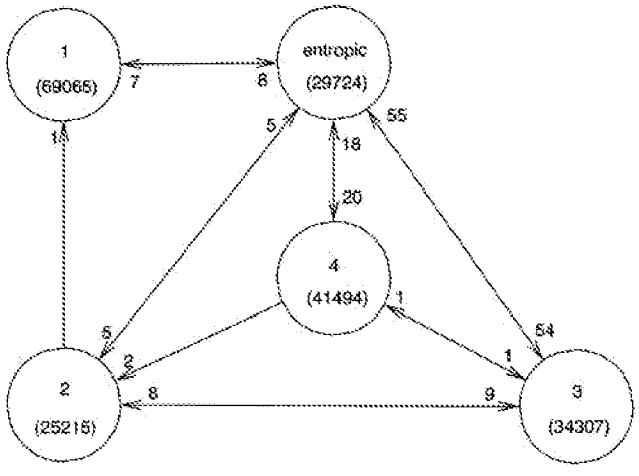

Fig. 5. Transition matrix between major basins represented as a simplified network. The numbers in brackets show the number of times system visited the basin (which corresponds to the partition function of the basin) and the numbers on the edges show the number of direct transitions in each direction between the basins. We note that the number of direct transitions is different from those appearing in Fig. 2, which is based on all transitions (direct and indirect) between the basins; the latter measures the overall flow, and therefore the barriers, between the basins (4).

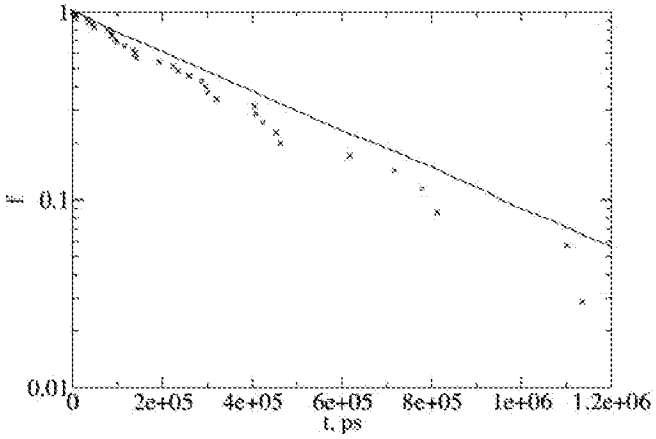

Fig. 6. Cumulative distribution (f) of the first passage times for reaching the native structure, starting form the completely extended structure; $f(t) = \int_t^\infty p(\tau)d\tau$, where p is the probability distribution of the first passage time. Crosses are for the MD simulation (35 events, the mean first passage time is $\approx$350 ns) and the line is for MC simulation with {T} transition matrix (10,000 events, the mean first passage time is $\approx$422 ns) (see *Methods*).

sampling of the whole configurational space (28). It also should be noted that the fastest pathways are not necessarily the most probable ones, as has been pointed out by Fersht (9) and Paci *et al.* (30).

Use of a disconnectivity graph to represent the FES reveals its inherent complexity. Most importantly, the denatured basin has a number of deep subbasins with low enthalpy and low entropy, which are not evident in projected surfaces. In fact, several of these basins have an energy below that of the native basin, which is stabilized by its higher entropy. This result contrasts with the standard funnel picture, which assumes a single large basin for the effective energy that directs folding to the native state. Although, as pointed out above, the effective energy surface is of interest, it is the FES that determines the folding behavior. Since it is very complex in the present case, relative to 2D FES projections, it is necessary to be cautious about the usual interpretation of the latter. That the available experimental data can be parametrized in terms of a 1D Kramers-like model (11) may tell us more about the limitations of the measurements than the underlying phenomenon; i.e., as long as the transitions among the different minima of the denatured basins are fast, their signature does not appear in the folding kinetics. A simple model can then describe the kinetics, but one must realize that the information obtained about the denatured FES is limited. It remains to be determined whether the deep nonnative basins found here have been eliminated in proteins by evolution, as has been suggested (ref. 7 but see also refs. 31 and 32), or whether they do exist and have not been observed, as yet.

We thank Aaron Dinner for helpful discussions and assistance in the comparison with the results in ref. 13 and William Eaton for helpful comments on the manuscript. Partial support for the research done at Harvard was provided by National Institutes of Health Grant GM30804.

1. Becker, O. M. & Karplus, M. (1997) *J. Chem. Phys.* **106,** 1495–1517.
2. Wales, D. J., Miller, M. A. & Walsh, T. R. (1998) *Nature* **394,** 758–760.
3. Levy, Y. & Becker, O. M. (1998) *Phys. Rev. Lett.* **81,** 1126–1129.
4. Krivov, S. V. & Karplus, M. (2002) *J. Chem. Phys.* **117,** 10894–10903.
5. Wales, D. J. (2003) *Energy Landscapes* (Cambridge Univ. Press, Cambridge, U.K.).
6. Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. (1995) *Science* **267,** 1619–1620.
7. Onuchic, J. N. & Wolynes, P. G. (2004) *Curr. Opin. Struct. Biol.* **14,** 70–75.
8. Lazaridis, T., Archontis, G. & Karplus, M. (1995) *Adv. Protein Chem.* **47,** 231–306.
9. Fersht, A. R. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 14122–14125.
10. Gronenborn, A. M., Filpula, D. R., Essig, N. Z., Achari, A., Whitlow, M., Wingfield, P. T. & Clore, G. M. (1991) *Science* **253,** 657–661.
11. Munoz, V., Thompson, P. A., Hofrichter, J. A. & Eaton, W. A. (1997) *Nature* **390,** 196–199.
12. Honda, S., Kobayashi, N. & Munekata, E. (2000) *J. Mol. Biol.* **295,** 269–278.
13. Dinner, A. R., Lazaridis, T. & Karplus, M. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 9068–9073.
14. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swami-nathan, S. & Karplus, M. (1983) *J. Comp. Chem.* **4,** 187–217.
15. Neria, E., Fisher, S. & Karplus, M. (1996) *J. Chem. Phys.* **105,** 1902–1921.
16. Lazaridis, T. & Karplus, M. (1999) *Proteins* **35,** 133–152.
17. Zhou, R. H. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 14931–14936.
18. Ford, L. R. & Fulkerson, D. R. (1956) *Can. J. Math.* **8,** 399–404.
19. Gomory, R. E. & Hu, T. C. (1961) *SIAM J. Appl. Math.* **9,** 551–570.
20. Zhou, R. H. (2003) *Proteins* **53,** 148–161.
21. Pande, V. S. & Rokhsar, D. S. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 9062–9067.
22. Bolhuis, P. G. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 12129–12134.
23. Wei, G., Mousseau, N. & Derreumaux, P. (2004) *Proteins* **56,** 464–474.
24. Dill, K. A. & Chan, H. S. (1997) *Nat. Struct. Biol.* **4,** 10–19.
25. Boczko, E. M. & Brooks, C. L., III (1995) *Science* **269,** 393–396.
26. Levy, Y., Jortner, J. & Becker, O. M. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 2188–2193.
27. Tang, Y. F., Rigotti, D. J., Fairman, R. & Raleigh, D. P. (2004) *Biochemistry* **43,** 3264–3272.
28. Evans, D. A. & Wales, D. J. (2004) *J. Chem. Phys.* **121,** 1080–1090.
29. Wales, D. J. (2002) *Mol. Phys.* **100,** 3285–3305.
30. Paci, E., Cavalli, A., Vendruscolo, M. & Caflisch, A. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 8217–8222.
31. Taverna, D. M. & Goldstein, R. A. (2002) *J. Mol. Biol.* **315,** 479–482.
32. Go, N. (1999) in *Old and New Views of Protein Folding*, eds. Kuwajima, K. & Arai, M. (Elsevier, Amsterdam), pp. 97–108.
33. Garcia-Viloca, M., Gao, J., Karplus, M. & Truhlar, D. G. (2004) *Science* **303,** 186–195.